Choosing the best machine translation system to translate a sentence by using only source-language information*

Felipe Sánchez-Martínez
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071, Alacant, Spain
fsanchez@dlsi.ua.es

Abstract

This paper describes a novel approach aimed to identify *a priori* which subset of machine translation (MT) systems among a known set will produce the most reliable translations for a given source-language (SL) sentence. We aim to select this subset of MT systems by using only information extracted from the SL sentence to be translated, and without access to the inner workings of the MT systems being used. A system able to select in advance, without translating, that subset of MT systems will allow multi-engine MT systems to save computing resources and focus on the combination of the output of the best MT systems. The selection of the best MT systems is done by extracting a set of features from each SL sentence and then using maximum entropy classifiers trained over a set of parallel sentences. Preliminary experiments on two European language pairs show a small, non-statistical significant improvement.*

1 Introduction

Machine translation (MT) has become a viable technology that helps individuals in assimilation —to get the gist of a text written in a language the reader does not understand— and dissemination —to produce a draft translation to be post-edited for publication— tasks. However, none of the different approaches to MT, whether statistical (Koehn, 2010), example-based (Carl and Way, 2003), rule-based (Hutchins and Somers, 1992) or hybrid (Thurmair, 2009), always provide the best results. This is why some researchers have investigated the development of multi-engine MT (MEMT) systems (Eisele, 2005; Macherey and Och, 2007; Du et al., 2009; Du et al., 2010) aimed to provide translations of higher quality than those produced by the isolated MT systems in which they are based on.

MEMT systems can be classified according to how they work. On one hand, we find systems that combine the translations provided by several MT systems into a consensus translation (Bangalore et al., 2001; Bangalore et al., 2002; Matusov et al., 2006; Heafield et al., 2009; Du et al., 2009; Du et al., 2010); the output of these MEMT systems may differ from those provided by the individual MT systems they are based on. On the other hand, we have systems that decide which translation, among all the translations computed by the MT systems they are based on, is the most appropriate one (Nomoto, 2004; Zwarts and Dras, 2008) and output this translation without changing it in any way. In-between, we find the MEMT systems that build a consensus translation from a reduced set of translations, i.e. systems that first chose the subset with the most promising translations, and then combine these translations to produce a single output (Macherey and Och, 2007).

Even though MEMT systems that select the most promising translation and those that work on a reduced subset of translations do not use all the translations computed by all the MT system, both kinds of MEMT systems need to translate the input source-language (SL) sentence as many times as different MT systems they use. This fact makes it difficult to integrate MEMT systems in environments where response time and required resources (mainly amount of memory and computing speed) are constrained. In addition, this also forces MEMT systems to keep the amount of MT systems they

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 97–104
Leuven, Belgium, May 2011
* See erratum at the end of the paper.

use to a minimum in order to keep the amount of needed resources low.

In this paper we describe a novel approach aimed to identify *a priori* which subset of MT systems among a known set will produce the most reliable translations for a given SL sentence. A system able to select in advance, without translating, that subset of MT systems from a known set of MT systems will allow MEMT systems to save computing resources and focus on the combination of the output of the best MT systems. At the same time, such a tool will allow the number of MT systems in which current MEMT systems are based to be increased.

The selection of the best MT systems is done by extracting a set of features from each SL sentence and then using maximum entropy classifiers trained over a set of parallel sentences. During training the source sentences in the training parallel corpus are automatically translated with the different MT systems being considered, and then the target sentences are evaluated against the reference translations in the training parallel corpus. To automatically determine the MT system producing the best translation during training we have tried several MT evaluation measures at the sentence level.

The rest of the paper is organised as follows. Next section presents the SL features used to discriminate between the different MT systems, and explains the training procedure and the way in which the classifiers are used for the task at hand. Section 3 then describes the experiments conducted, whereas results are discussed in Section 4. The paper ends with some concluding remarks and plans for future work.

2 System selection as a classification problem

We aim to select the subset of MT systems that will produce the best translations by using only information extracted from the source sentence to translate, without access to the inner workings of the MT systems being used. To achieve this goal we have used binary maximum entropy classifiers (see below) and tried several features, some of which needs the input sentence to be parsed by means of a statistical parser (see Section 3 to know about the parser we have used),[1] while the others can be easily obtained from the SL sentence. Note that some of the (SL) features we have used have also been used in combination with other features for sentence-level confidence estimation (Blatz et al., 2003; Quirk, 2004; Specia et al., 2009), a related task aimed at assessing the correctness of a translation. A description of the features we have tried follows:

- maximum depth of the parse tree [`gmaxd`],
- mean depth of the parse tree [`gmeand`],
- joint likelihood of the parse tree t and the words w in the sentence, i.e. $p(t, w)$ [`gjl`],
- likelihood of the parse tree given the words, i.e. $p(t|w)$ [`gcl`],
- sentence likelihood as provided by the model used to parse the sentence, i.e. summing out all possible parse trees [`gsentl`],
- maximum number of child nodes per node found in the parse tree [`gmaxc`],
- mean number of child nodes per node [`gmeanc`],
- number of internal nodes [`gint`],
- number of words whose mean *shift* (see below) is greater than a given threshold (values used: 1, 2, 3, 4, 5) [`smean`],
- number of words whose variance over the shift is greater than a given threshold (values used: 2, 4, 6, 8, 10) [`svar`],
- number of words whose mean *fertility*, i.e. the mean number of target words to which a source word is aligned, is greater than a given threshold (values used: 0.25, 0.5, 0.75, 1, 1.25, 1.50, 1.75, 2) [`fmean`],
- number of words whose variance over the fertility is greater than a given threshold (values used: 0.25, 0.5, 0.75, 1, 1.25, 1.50, 1.75, 2) [`fvar`],
- sentence length in words [`len`],
- number of words not appearing in the corpora used to trained the corpus-based MT system used [`unk`], and

[1] It may be argued that parsing a sentence may be as time consuming as translating it; however, in MEMT a sentence is translated several times, and thus avoiding to perform such translations, even by using computationally expensive procedures such as parsing, helps saving computational resources because each sentence is parsed only once.

- likelihood of the sentence as provided by an n-gram language model trained on a SL corpus [slm].

The shift of a source word at position i is defined as $\text{abs}(j - i)$, where j is the position of the first target word to which that source word is aligned. In the experiments we computed the mean and variance of both the shift and the fertility from a parallel corpus by computing word alignments in the usual way, i.e. by running GIZA++ (Och and Ney, 2003) in both translation directions and then symmetrising both sets of alignments through the "grow-diag-final-and" heuristic (Koehn et al., 2003) implemented in MOSES (Koehn et al., 2007). We then use these pre-computed values when obtaining the features of an input sentence.

The features obtained from the parse tree of the sentence try to describe the sentence in terms of the complexity of its structure. The features related to the shift and the fertility of the words to be translated are intended to describe the sentence in terms of the complexity of its words. The rest of features —sentence length, likelihood of the sentence to be translated and number of words not appearing in the parallel corpora used to train the corpus-based MT systems— might be helpful to discriminate between the rule-based MT systems and the corpus-based ones.

To find the set of relevant features we have used the chi-square method (Liu and Setiono, 1995) that evaluates features individually. We ranked all the features according to their chi-squared statistic (DeGroot and Schervish, 2002, Sec. 7.2) with respect to the classes and select the first N features in the ranking. To determine the best value of N we evaluated the translation performance achieved on a development corpus with all possible values of N.

Training. For each MT system used we have trained a maximum entropy model (Berger et al., 1996) that will allow our system to compute for an input sentence the probability of that sentence being best translated by each system. In order to train these classifiers, and for each different evaluation measure we have tried, each parallel sentence in the training corpus is preprocessed as follows:

1. the SL sentence is translated into the TL through all the MT systems;

2. each translation is evaluated against the reference translation in the training parallel corpus;

3. all the machine translated sentences are ranked according to the evaluation scores obtained, and the subset of MT system producing the best translation are determined; note that it may happen that several MT systems produce the same translation, or that several machine translated sentences are assigned the same score.

After this preprocessing, the corpus of instances from which the binary classifier associated to an MT system is trained consist of as many instances as parallel sentences in the training corpus. Each instance in this corpus is classified as belonging to the class of that MT system if it appears in the subset of MT systems producing the translation(s) leading with the best evaluation score.

System selection. When a SL sentence is to be translated, first the sentence is parsed, and the features described above are extracted; then, the probability of each MT system being the best system to translate that sentence is estimated by means of the different maximum entropy models. The systems finally selected to translate the input sentence are the ones with the highest probabilities. In this papers we have tested this approach by selecting only a single MT system, the one with the highest probability.

3 Experimental settings and resources

We have tested our approach in the translation of English and French texts into Spanish. The systems we have used are: the shallow-transfer rule-based MT system APERTIUM (Forcada et al., 2011),[2] the rule-based MT system SYSTRAN (Surcin et al., 2007),[3] the phrase-based statistical MT system MOSES (Koehn et al., 2007),[4] the MOSES-CHART hierarchical phrase-based MT (Chiang, 2007) system, and the hybrid example-based–statistical MT system CUNEI (Phillips and Brown, 2009).[5]

The three corpus-based systems, namely MOSES, MOSES-CHART and CUNEI, were trained using the data set released as part of the WMT10 shared translation task.[6] The corpora used to train and evaluate the five binary maximum entropy classifiers were

[2] http://www.apertium.org

[3] We have used the version of Systran provided by Yahoo! Babelfish: http://babelfish.yahoo.com

[4] http://www.statmt.org/moses/

[5] http://www.cunei.org

[6] http://www.statmt.org/wmt10/training-parallel.tgz

Pair	Corpus	Num. sent.	Num. words
en-es	Training	98,480	en: 2,996,310; es: 3,420,636
	Development	1,984	en: 49,003; es: 57,162
	Test	1,985	en: 55,168; es: 65,396
fr-es	Training	99,022	fr: 3,513,404; es: 3,449,999
	Development	1,987	fr: 60,352; es: 59,551
	Test	1,982	fr: 64,392; es: 64,440

Table 1: Number of sentences and words in the corpora used to train and evaluate our MT system(s) selection approach.

extracted from the corpus of the United Nations that is also distributed as part of the WMT10 shared translation task. The French–Spanish parallel corpus was obtained from the English–French and the English–Spanish parallel corpora by pairing French and Spanish sentences having as translation the same English sentence.[7] After removing duplicated sentences and sentences longer than 200 words, we used the first 2,000 sentences for development, the second 2,000 sentences for testing, and the next 100,000 sentences for training. Note that some sentences in these corpora could not be parsed with the parser we have used (see below) and, therefore, they were removed before running the experiments. Table 1 provides detailed information about these corpora and the number of sentences finally used in the experiments.

To parse the input SL sentences we used the Berkeley Parser (Petrov et al., 2006; Petrov and Klein, 2007) together with the parsing models available for English and French from the parser website.[8] To compute the likelihood of the SL sentences we used a 5-gram language model trained by means of the IRSTLM language modelling toolkit[9] (Federico et al., 2008) by using the SL corpora distributed as part of the WMT10 shared translation task. Variance and mean shifts and fertilities were calculated on the same corpora used to train the corpus-based MT systems.

After translating the SL sentences in the training corpora through all the MT systems being considered, we used the ASIYA evaluation toolkit[10] (Giménez and Màrquez, 2010) to evaluate, at the sentence level, the translation provided by each MT system against the TL reference in the training parallel corpora. For that we used the precision-oriented measure BLEU (Papineni et al., 2002), two edit distance-based measures, PER and TER (Snover et al., 2006); and METEOR (Lavie and Agarwal, 2007), a measure aimed at balancing precision and recall that considers stemming and, only for some languages, synonymy lookup using WordNet. In our experiments we only used stemming when computing the lexical similarity of two words.

To train and test the five binary maximum entropy classifiers we used the WEKA machine learning toolkit (Witten and Frank, 2005) with default parameters; the class implementing the maximum entropy classifier is `weka.classifiers.functions.Logistic`. The class implementing the chi square method we used to select the set of relevant features on a development corpus is `weka.attributeSelection.ChiSquaredAttributeEval`.

With respect to the instances used to train the five binary maximum entropy classifiers and how many times an instance happens to belong to more than a class (MT system), Table 2 reports the percentage of sentences in the training corpora for which the translation or translations being assigned the best evaluation score are produced by M different MT systems. Recall that M may be greater than one because more than an MT system may produce the same translation or because more than a machine translated sentence may be assigned the same evaluation score. It is worth noting that the percentage of sentence for which the output of more than an MT system gets the highest score is larger in the case of TER and PER than in the case of the other two evaluation measures.

4 Results and discussion

Table 3 reports, for the two language pairs we have tried, the translation performance, as measured by different MT evaluation measures, achieved by the

[7] Original corpora can be downloaded from `http://www.statmt.org/wmt10/un.en-fr.tgz` and `http://www.statmt.org/wmt10/un.en-es.tgz`

[8] `http://code.google.com/p/berkeleyparser/`

[9] `http://hlt.fbk.eu/en/irstlm`

[10] `http://www.lsi.upc.edu/~nlp/Asiya/`

Pair	Measure	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
en-es	BLEU	82.8%	9.9%	5.6%	0.8%	0.9%
	PER	58.7%	23.6%	12.5%	3.5%	1.7%
	TER	62.1%	22.3%	11.1%	2.9%	1.6%
	METEOR	83.5%	9.3%	5.4%	0.7%	1.1%
fr-es	BLEU	74.4%	12.8%	6.4%	3.3%	3.1%
	PER	51.6%	21.9%	13.7%	7.2%	5.6%
	TER	52.6%	22.1%	13.2%	6.7%	5.4%
	METEOR	74.0%	11.9%	5.9%	3.0%	5.2%

Table 2: Percentage of sentences in the training corpora for which the best evaluation score is assigned to the translation or translations produced by M different MT systems.

Pair	Configuration	BLEU	PER	TER	METEOR
en-es	Best system	0.3481 (M)	0.3581 (MC)	0.4851 (M)	0.2745 (C)
	System selection	0.3529 (11)	0.3582 (3)	0.4838 (8)	0.2762 (13)
	Oracle	0.3905	0.3299	0.4409	0.2965
fr-es	Best system	0.3146 (C)	0.4128 (C)	0.5880 (C)	0.2281 (C)
	System selection	0.3192 (19)	0.4109 (16)	0.5861 (16)	0.2286 (22)
	Oracle	0.3467	0.3913	0.5548	0.2389

Table 3: Performance achieved by the best MT system, by the systems selected through our approach (system selection), and by the combination of translations providing the best possible performance (oracle). The system achieving the best performance at the corpus level and the number of features used by our approach are reported between brackets. M stands for MOSES, MC for MOSES-CHART, and C for CUNEI.

best MT system at the corpus level (reported between brackets), the performance achieved by our approach, and that of the oracle, i.e the best possible performance. The latter was calculated by translating all the SL sentences in the test corpus through all the MT system being used, and then selecting for each sentence the translation getting the best evaluation score. The number of features used by our approach after feature selection is reported between brackets.

Results in Table 3 show that our method very slightly improves the performance achieved by the best MT system for both language pairs, although this small improvement is larger in the case of English–Spanish. 95% confidence intervals computed by bootstrap resampling (Koehn, 2004) show a large overlapping between the performance achieved by the best system and that of our system selection approach. Note that no overlapping occurs between the confidence intervals of the best system and that of the oracle. It is worth noting that on the development corpus the improvement was larger for fr-es than for en-es, although still very small to be statistically significant.

A manual inspection of the first 500 sentences in the en-es test corpus together with their automatic translations show that most of the times the MT systems produce translations of similar quality, and therefore it is hard to chose one of them as the best translation. For the first 500 sentences in the en-es test corpus we ranked the translations provided by the different MT systems we have used, without access to the reference translation, and found out that the difference between the BLEU score achieved by the best performing MT system for the first 500 sentences of the en-es test corpus, i.e. MOSES, (0.3926) and that of the best translation manually selected (0.3928) is even lower than the one obtained through our approach. This may be explained by the fact that the three corpus-based systems we have used were trained on the same parallel corpora and also because of the homogeneity of the corpora we have used for training and testing.

With respect to the number of times each system is chosen by our approach when translating the test corpora, Table 4 reports the percentage of time this happens for each system and MT evaluation measure. Note that when the en-es system selection is trained using PER, most of the times it

Pair	Measure	M	MC	C	A	S
en-es	BLEU	32.9%	51.1%	2.6%	0.1%	13.3%
	PER	2.9%	95.8%	0.0%	0.0%	1.3%
	TER	53.6%	36.0%	5.5%	0.0%	4.9%
	METEOR	28.8%	18.5%	41.8%	0.0%	10.9%
fr-es	BLEU	0.2%	42.5%	38.1%	0.0%	19.2%
	PER	0.0%	28.4%	59.8%	0.0%	11.8%
	TER	0.2%	36.7%	53.7%	0.0%	9.4%
	METEOR	0.0%	26.6%	63.2%	0.0%	10.2%

Table 4: Percentage of times each systems is chosen when translating the test corpora. M stands for MOSES, MC for MOSES-CHART, C for CUNEI, A for APERTIUM, and S for SYSTRAN.

chooses MOSES-CHART; it may be concluded that the reduced number of features chosen by the feature selection method on the development corpus for this language pair and evaluation measure does not allow the system to discriminate between the different MT systems.

Finally, the features that happen to be relevant with the majority of evaluation measures are (see Section 2 for a description of each one)

- for `en-es`: `gmaxd`, `gmeand`, `gcl`, `gsentl`, `smean` for thresholds 1 and 2, and `svar` for thresholds 2, 4 and 6; and
- for `fr-es`: `len`, `gcl`, `gsentl`, `gint`, `smean` for thresholds 1 and 2, `svar` for thresholds 2, 4, 6, and 10, `fmean` for thresholds 0.25, 0.5 and 0.75, and `slm`.

5 Concluding remarks

In this paper we have presented a novel approach aimed to select the subset of MT systems, among a known set of systems, that will produce the most reliable translations for a given sentence by using only information extracted from that sentence. Preliminary experiments in the translation of English and French texts into Spanish shows a small, non-statistically-significant improvement compared to the translation provided by the MT system performing best on the whole test corpus. In addition, a manual selection of the best MT system on a per-sentence basis shows that it is hard to perform such a selection because most of the sentences are translated similarly with most of the MT systems.

As a future work we plan to try different configurations of WEKA as well as use a development corpus to tune the trained classifiers. We also plan to incorporate new features, use MT systems trained on different corpora, use corpora with sentences coming from different sources, and evaluate the translation performance when a fixed number of MT systems are selected through our approach and then their translations are combined using MANY (Barrault, 2010).

Acknowledgements

We thank Sergio Ortiz-Rojas for his help and ideas, Inmaculada Ruiz López for performing the manual ranking of the first 500 sentences in the English–Spanish test corpus, and Mikel L. Forcada and Francis M. Tyers for their help with the manuscript. Work funded by the European Association for Machine Translation through its 2010 sponsorship of activities program and by the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01.

References

Bangalore, S., G. Bordel, and G. Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 351–354.

Bangalore, S., V. Murdock, and G. Riccardi. 2002. Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of 19th International Conference on Computational Linguistics*, pages 1–7, Taipei, Taiwan.

Barrault, L. 2010. MANY: open source machine translation system combination. *Prague Bulletin of Mathematical Linguistics*, (93):147–155. Fourth Machine Translation Marathon. Dublin, Ireland.

Berger, A. L., S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Blatz, J., E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for machine translation. Technical report, Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

Carl, M. and A. Way, editors. 2003. *Recent Advances in Example-Based Machine Translation*, volume 21. Springer.

Chiang, D. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

DeGroot, M. H. and M. J. Schervish. 2002. *Probability and Statistics*. Addison-Wesley, third edition.

Du, J., Y. Ma, and A. Way. 2009. Source-side context-informed hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the Twelfth Machine Translation Summit*, pages 230–237, Ottawa, ON, Canada.

Du, J., P. Pecina, and A. Way. 2010. An augmented three-pass system combination framework: DCU combination system for WMT 2010. In *Proceedings of the Fifth ACL Workshop on Statistical Machine Translation*, pages 271–276, Uppsala, Sweden.

Eisele, A. 2005. First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 155–158, Ann Arbor, MI, USA.

Federico, M, N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*, pages 1618–1621, Melbourne, Australia.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*. Special Issue on Free/Open-Source Machine Translation (In press).

Giménez, J. and L. Màrquez. 2010. Asiya: an open toolkit for automatic machine translation (meta-)evaluation. *Prague Bulletin of Mathematical Linguistics*, (94):77–86. Fifth Machine Translation Marathon. Le Mans, France.

Heafield, K., G. Hanneman, and A. Lavie. 2009. Machine translation system combination with flexible word ordering. In *Proceedings of the Fourth ACL Workshop on Statistical Machine Translation*, pages 56–60, Suntec, Singapore.

Hutchins, W. J. and H. L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Morristown, NJ, USA.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press.

Lavie, A. and A. Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228—-231, Prague, Czech Republic.

Liu, H. and R. Setiono. 1995. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, pages 388–391.

Macherey, W. and F. J. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 986–995, Prague, Czech Republic.

Matusov, E., N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40, Trento, Italy.

Nomoto, T. 2004. Multi-engine machine translation with voted language model. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 494–501, Barcelona, Spain.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Petrov, S. and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 404–411, Rochester, NY, USA.

Petrov, S., L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia.

Phillips, A. B. and R. D. Brown. 2009. Cunei machine translation platform: system description. In *Proceedings of the Third Workshop on Example-Based Machine Translation*, pages 29–36, Dublin, Ireland.

Quirk, C. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of the The fourth international conference on Language Resources and Evaluation*, pages 525–828, Lisbon, Portugal.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

Specia, L., M. Turchi, Z. Wang, J. Shawe-Taylor, and C. Saunders. 2009. Improving the confidence of machine translation quality estimates. In *Proceedings of Twelfth Machine Translation Summit*, pages 136–143, Ottawa, Canada.

Surcin, S., E. Lange, and J. Senellart. 2007. Rapid development of new language pairs at SYSTRAN. In *Proceedings of the Eleventh MT Summit*, pages 443–449, Copenhagen, Denmark.

Thurmair, G. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceedings of the Twelfth Machine Translation Summit*, pages 340–347, Ottawa, ON, Canada.

Witten, I. H. and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Elsevier, second edition.

Zwarts, S. and M. Dras. 2008. Choosing the right translation: a syntactically informed classification approach. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1153–1160, Manchester, UK.

Erratum

Statistical significant tests performed by pair bootstrap resampling (Koehn, 2004) show that the difference in performance between the system performing best at the document level and that of the system selection approach described in this paper is statistically significant with p=0.05 for all the automatic MT evaluation metrics we have used, with the exception of the METEOR scores obtained for the French-Spanish language pair.